Increasing Object-Level Reconstruction Quality in Single-Image 3D Scene Reconstruction

Anna Ribic Antonio Oroz Meikel Kokowski Franz Srambical
Technical University of Munich
{firstname}.{lastname}@tum.de

Abstract

1. Introduction

While humans can easily infer the 3D structure as well as the complete (panoptic) semantics of a scene from a single image, this task has been a longstanding challenge in the field of computer vision. The task fundamentally prerequisites learning a strong prior of the 3D world. Traditional methods have made significant strides, from generating geometrically coherent structures [**?** **?**] to learning different instance semantics [**?** **?** **?**]. More recent approaches directly learn the 3D panoptic semantics as a whole [**?** **?**], yet they fall short in capturing the intricate details and nuances at the object level. This paper introduces a novel approach to bridge this gap by integrating a specialized object-level model into the reconstruction process, thereby leveraging the specialized model's object-priors.

Our approach models panoptic 3D reconstruction as a two-stage problem. We first use the model of **?**] to create an initial reconstruction. Then, we leverage the instance masks to extract the object geometries out of the reconstructed scene. We input each of the extracted objects along with cropped images from the scene and text labels into a diffusion model [**?**] to refine the rough object-level geometries. Finally, we integrate the refined object geometries back into the initial scene reconstruction to obtain a complete and refined panoptic 3D scene reconstruction.

In summary, our main contributions are as follows:

- We propose a novel approach to panoptic 3D reconstruction involving an inference pipeline that leverages object-level reconstruction models to refine the output of a 3D scene reconstruction backbone.
- We qualitatively demonstrate the effectiveness of our approach on the 3D-Front [**?**] dataset, showing significant improvements over the state-of-the-art.
- We show that fine-tuning SDFusion [**?**] on the input scene's object distribution (in our case the 3D-Future dataset [**?**]) significantly improves the quality of the refined objects.
- We propose *weighted masking*, a novel technique to integrate masking uncertainty into the object-level reconstruction process.
- We introduce a conceptually simple yet effective method for shape alignment, which outperforms rigid alignment methods in our experiments.
- We openly release our model code, training and inference pipelines, as well as our newly constructed variation of the 3D-Front dataset to facilitate future research in the field.

2. Related Work

2D panoptic segmentation 2D panoptic segmentation merges semantic and instance segmentation, providing detailed pixel-level parsing of images, capturing both general categories (semantic segmentation) and individual object identities (instance segmentation) [**?**]. Since the original task formulation by **?**], a number of works have been proposed to solve the task [**?** **?** **?** **?** **?** **?** **?** **?** **?** **?** **?** **?** **?**], while more recent approaches [**?**] try to unify image segmentation in its entirety.

Single-view 3D reconstruction The work by **?**] was the first notable attempt at reconstructing 3D scenes from unordered photo collections. Since then, the field of image-based 3D reconstruction has seen a number of advancements, culminating in the task of single-view 3D reconstruction [**?** **?** **?** **?** **?** **?** **?**].

Shape priors **?**] note that the task of single-view 3D reconstruction is non-deterministic, as there are many 3D shapes that can explain a given single-view input, and propose to use shape priors to shape the solution space such that the reconstructed shapes are realistic, but not necessarily the ground truth.

3D scene understanding The task of 3D scene understanding and panoptic reconstruction is analogous to its 2D

counterpart and aims to infer the 3D structure and semantics of a scene, including the 3D layout, object instances, and their 3D shapes from images [**?**] or noisy geometry [**? ?**]. **?**] propose a method – henceforth called *Panoptic 3D* – to jointly solve the tasks of 3D scene understanding and single-view 3D reconstruction by lifting features produced by a 2D backbone into a 3D volume of the camera frustrum, and jointly optimizing for geometric reconstruction as well as semantic and instance segmentation.

Modality-conditioned shape generation 3D generative models represent objects in a variety of modalities, including point clouds [**? ?**], occupancy grids [**?**], meshes [**?**], and signed distance functions [**?**]. Furthermore, these models can also be distinguished by the type of input they take, such as incomplete shapes [**?**], images [**?**], text [**? ?**], or other modalities [**?**]. Notably, **?**] propose *SDFusion*, a 3D object reconstruction method conditioned on images, text and geometrical input.

Datasets Research in 3D panoptic reconstruction relies heavily on datasets to train and evaluate algorithms. Notable datasets in this domain include ScanNet (cite here) and Replica (cite here), which provide rich annotations for scene understanding tasks.

Among these datasets, the 3D Front dataset stands out for its comprehensive coverage of indoor scenes. Created by Li et al. (cite here), 3D Front offers detailed geometric reconstructions, semantic segmentation, and instance segmentation annotations for various indoor environments, including living rooms, kitchens, and bedrooms. The synthetic 3D dataset contains 6,801 mid-size apartments with 18,797 rooms populated by 3D shapes from the 3D-Future (cite here) dataset. The dataset's high-quality data acquisition process ensures accurate representations, making it a valuable resource for advancing research in 3D panoptic reconstruction.

3. Method

4. Conclusion